October 29, 2019

Federal Communications Commission
Office of Disability Rights
445 12th Street SW
Washington, DC 20554

RE: Petition for Declaratory Ruling and/or Rulemaking on Live Closed Captioning Quality Metrics and the Use of Automatic Speech Recognition Technologies

To Whom It May Concern:

As a global leader and technical innovator of Automated Speech Recognition (ASR) technologies, IBM appreciates the opportunity to respond to this petition. IBM has a 100-year history of technical innovation in support of people with disabilities. This innovation includes everything from Braille typewriters in the 1960's to the cognitive voice navigation systems used to support accessible route guidance via mobile applications today. Our ASR technologies follow generations of research, award-winning innovation, and decades of commitment to building technology that enhances the human experience.

With emerging applications of artificial intelligence (AI) in ASR technology, IBM is committed to promoting trust in these technologies by introducing them into the world in a responsible manner. We believe the purpose of AI is to augment—not replace— human intelligence and our technology is designed to enhance and extend human capability and potential. In addition, we believe the benefits of the AI era should touch the many, not just the elite few. These beliefs guide IBM's development of ASR technology that will ultimately lead to an improved final experience for the Deaf and hard-of-hearing community.

As such, we write to express our support for other commentators who shared three key concerns in regard to the petition:

- a declaratory ruling is premature given the nascency of the ASR technology;
- innovation in ASR technology that enhances the experience of those using captioning, such as the Deaf and hard-of-hearing community, should be encouraged; and
- caption quality can best be achieved by showcasing fidelity to existing quality standards metrics, which requires incorporating regular direct feedback from the Deaf and hard-of-hearing community.

We appreciate the opportunity to share our unique perspective to these issues in our response below. As the Agency weighs future actions, we welcome continued engagement. For any questions, please contact **me at dkulczar@us.ibm.com.**

Sincerely,

David Kulczar
Program Director, Offering Management, IBM Watson Media Solutions

Detailed Response

Many of IBM's concerns with regards to the petition for a declaratory rulemaking are shared by other commentators in the docket. In particular, we agree with NAB that

> a declaratory ruling is both premature and unnecessary given the nascency of ASR and the negative impact that special certifications or other new rules could have on its development. It is critical that the Commission preserve and promote all options for creating captions, old and new. Moreover, the existing best practices sufficiently align with the steps that ASR users already take to improve caption quality.[1]

Additionally, we offer the following input for consideration by the Agency on both points of the petition: Live Closed Captioning Quality Metrics and Use of Automatic Speech Recognition Technologies:

Caption Quality Standards

IBM has been working on speech-related technologies for decades to extend human capability that aims to enhance the experience of the Deaf and hard-of-hearing community. When this work started more than 40 years ago, we were able to capture only a few words successfully. However, over the past decade, the rapid growth of machine learning technologies, improved audio techniques, and better network capabilities have accelerated the quality of ASR technologies significantly, where they now rival traditional captioning methods. As these technologies continue to mature, so too will the demand for individuals with captioning skillsets to train and support AI that can help drive an improved ASR service.

Specifically, if we look directly at the Federal Communications Commission's (FCC) four caption quality standards – accuracy, synchronicity, completeness, and appropriate placement – ASR developers directly improve on traditional methods in three of these four areas today (the quality of "appropriate placement" remains largely the purview of broadcasters, outside the direct control of caption technology developers and suppliers).

1. **Accuracy.** Our measurements have shown that within one month of station training, ASR technology achieves accuracy levels on par with, or better than, traditional services. Although certain types of programming are currently a better fit for traditional captioning formats, such as content with large amounts of background noise or cross-talk, ASR is rapidly maturing in those dimensions.[2]

2. **Synchronicity.** With automated live broadcast captioning, we are able to consistently generate transmission delays that average less than two seconds, significantly improving

[1] *Opposition to Petition for Rulemaking*, National Association of Broadcasters, 15 Oct. 2019, p. 3, available at https://ecfsapi.fcc.gov/file/1015002782834/Caption%20Metrics%20Petition%20Opposition%2010-15-19.pdf.
[2] IBM's work with the US Open, for example, is one case study where automated captioning has shown great success with live sports formats.

on the latency delay of traditional human-operated caption delivery services. In terms of completeness, automated systems excel.

3. **Completeness.** Due to services such as built in redundancy, fault tolerance, and monitoring, ASR can ensure that broadcasts can be captured in their entirety. Where technical issues that may impact captioning availability do arise, such services are capable of immediately notifying providers so that corrective action can be taken.

ASR services also improve on traditional models in many other ways. In the event of breaking news that may impact public safety, for instance, it can be turned on and immediately begin providing the Deaf and hard-of-hearing community with important information via captioning services with zero delay. As this technology continues to mature, areas such as true video description will increasingly come within reach and significant research efforts are already underway.

Response to Petitioners

ASR technology can provide a crucial and socially beneficial service. IBM is committed to promoting trust in these technologies by developing them in a responsible manner. To build such trust, we constantly improve on accuracy, synchronicity, and completeness in caption quality standards, by, among other measures, incorporating regular direct feedback from the Deaf and hard-of-hearing community. This involves a constant feedback loop that relies on a flexible, adaptive, and dynamic environment conducive to ongoing experimentation and technological progress.

The petitioners note that one of their signatories–the Twenty-First Century Captioning Disability and Rehabilitation Research Project (Captioning DRRP)–has begun a project that, "[o]ver the coming years … will develop rigorous, scientifically sound, consumer-focused metrics for captioning quality and accompanying methods to conduct aggregate quality evaluations across the video programming ecosystem."[3] We welcome this initiative and associated efforts aimed at promoting mechanisms that may help promote future empirical work to improve the means by which caption quality standards are assessed. In addition, we also welcome an inquiry into the current state of the art in ASR and other closed captioning techniques and technologies in order to help further the goal of providing quality services to this community.

However, we believe an inquiry into the adoption and enforcement of new metrics would be premature at this time. Given the petitioner's admittedly nascent stage of research into objective quality metrics and the potential impact on ongoing innovation and maturation in ASR technology, we believe it would not be appropriate for the FCC to, as the petitioners request, "act with haste and issue expedited changed rules under Rule 1.412(c)." Rather, we believe the existing best practices promulgated under the *2014 Caption Quality Order* are appropriately suited to the application of ASR and have the added benefit of helping to foster an environment conducive to

[3] *Petition for Declaratory Ruling and/or Rulemaking on Live Closed Captioning Quality Metrics and the Use of Automatic Speech Recognition Technologies*, 31 July 2019, p. 12.

ongoing iterative improvements in this technology. Additionally, we believe that our ASR technology and quality assurance practices are in complete alignment with the expectations set out in the *2014 Caption Quality Order*, and that our technology is of at least a similar quality as the output provided by human captioners as facilitated by the caption best practices.

As the petitioners note, "[e]ven live human captioners can substantially omit or alter content to a degree that the original communicative intent of the audio track is no longer preserved."[4] Improving the experience of the Deaf and hard-of-hearing community, by addressing these human-centric captioning limitations, is one of many objectives that ASR technology seeks to realize. Yet, holding ASR to a higher standard than human captioners is not the way to achieve a higher quality experience for its users.

IBM has worked closely with our research team, existing clients, and several third-party companies to identify best practices related to accurately measuring captioning accuracy. Normally, this is done by employing a third-party to produce a baseline "ground truth" transcript on a small subset of captioned content; but this approach is limited. One of the primary limitations is that there is no objective measure or consistently reliable methodology for what one-hundred percent accuracy looks like in any caption transcript output. Two separate services will virtually always produce different results, whether due to different treatments in punctuation and grammar, imperfect sentence structures that are a regular feature of normal human speech, or a variety of other speech-based idiosyncrasies, everything from dialect and accent to vocal intonation and timbre.

IBM uses a few trusted vendors to test our ASR technology, however we also rely on technical capabilities such as Word Error Rate (WER) and F1 measurements to help predict accuracy. However, these methodologies tend to be more directional indicators of accuracy, as opposed to true scoring or measurement tools. By looking at WER trends over time, we can confidently point to improvement or degradation of a service, even if we cannot pinpoint a measurable percentage accuracy. Measurement tools and strategies, such as the research project noted by the petitioners, are still being researched and developed, and it is important for industry, vendors, community advocates, and the FCC to continue to monitor improvements in this area.

However, there is currently no truly objective way to either (1) produce a single accuracy score for ASR services or (2) a baseline measurement of an ideal or "perfect" caption transcription against which such a score could be compared. In addition to the problems mentioned above, any service's effectiveness will vary greatly depending on the specific attributes of the content; variations in speaker voice and grammar, cross talk, background noise, and volume issues can all directly impact overall accuracy for any service. Measurement strategies need to take into account and be mindful of these significant variables, which is only possible with quality standards that permit flexibility and experimentation.

We believe the existing models for measurement and reporting are effective mechanisms for showcasing fidelity to the quality standards metrics currently in use by the FCC. Captioning issues are currently rising to the attention of broadcasters through a number of channels and broadcasters

[4] *Id.* at 15.

and captioning vendors directly interfacing with those groups that rely on captioning is the single best way for captioning issues to come to the forefront and for action to be taken to improve captioning, regardless of delivery model.